scientific reports

Check for updates

OPEN

Dual knowledge-guided data augmentation for robust clinical prediction models

Sangwoo Moon[1], Peong Gang Park[2,3], Naye Choi[4], Ji Hyun Kim[5,6], Seon Hee Lim[7], Joo Hoon Lee[8], Min Ji Park[9], Hee Sun Baek[9], Min Hyun Cho[9], Keum Hwa Lee[2], Jae Il Shin[2,3], Kyoung Hee Han[10], Jeong Yeon Kim[11], Ji Yeon Song[7], Eun Mi Yang[12], Seong Heon Kim[4,6,13], Yo Han Ahn[4,6,13], Hee Gyung Kang[4,6,13] & Eujin Park[14]✉

Domain shift poses a critical barrier to adopting medical artificial intelligence (AI) models, a challenge that is particularly acute in data-scarce, single-source domain generalization (SSDG) settings. Conventional data-augmentation techniques, including Mixup and input masking, fail to leverage the rich structural information and expert knowledge inherent in clinical data, causing models to overfit by learning spurious correlations. To address this, we propose a dual knowledge-guided data augmentation framework that enhances model robustness by systematically embedding clinical expertise into the training process. It comprises two novel components: similarity-guided Mixup, which generates clinically plausible virtual samples by interpolating between patients with similar clinical profiles, and group-based masking, which simulates realistic missing-data patterns by concurrently masking clinically correlated features. We validated our framework on the multicenter KNOW-pedCKD cohort for pediatric chronic kidney disease, training it exclusively on a single-source domain and evaluating it on three unseen target domains. It demonstrated significant performance gains in recall, a metric of critical importance in clinical settings. This study demonstrates that embedding domain knowledge into data augmentation is a promising strategy for developing generalizable and trustworthy medical AI models capable of operating reliably across heterogeneous clinical environments. The codes are available at https://github.com/msw6468/dual_augmentation_public.

Keywords Domain shift, Data augmentation, Mixup, Masking, Clinical knowledge, Domain generalization, Chronic kidney disease

Advancements in artificial intelligence (AI) and machine learning have shown considerable potential, particularly within the medical domain. Clinical decision support systems derived from imaging signals and electronic health records data are emerging as core technologies[1–3]. These systems are pivotal for enhancing patient safety and treatment efficacy, providing diagnostic support to physicians and minimizing the risk of therapeutic errors.

Central to this technological progress are predictive models trained to learn meaningful patterns from complex datasets. These models are increasingly developed to predict disease risk, simulate progression, and forecast treatment responses using biological signals, laboratory results, imaging, and genetic data. AI holds particular promise for managing chronic diseases such as chronic kidney disease (CKD), a condition with high global prevalence that represents a major public health concern. The progression of CKD to end-stage kidney

[1]Department of Computer Science and Engineering, College of Engineering, Seoul National University, Seoul, Republic of Korea. [2]Division of Pediatric Nephrology, Severance Children's Hospital, Seoul, South Korea. [3]Department of Pediatrics, Yonsei University College of Medicine, Seoul, South Korea. [4]Department of Pediatrics, Seoul National University Children's Hospital, Seoul, South Korea. [5]Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, South Korea. [6]Department of Pediatrics, Seoul National University College of Medicine, Seoul, South Korea. [7]Department of Pediatrics, Pusan National University Children's Hospital, Busan, South Korea. [8]Department of Pediatrics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea. [9]Department of Pediatrics, Kyungpook National University, School of Medicine, Daegu, South Korea. [10]Department of Pediatrics, Jeju National University College of Medicine, Jeju, South Korea. [11]Department of Pediatrics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea. [12]Department of Pediatrics, Chonnam National University Hospital, Chonnam National University School of Medicine, Gwangju, South Korea. [13]Kidney Research Institute, Seoul National University Medical Research Center, Seoul, South Korea. [14]Department of Pediatrics, Korea University Guro Hospital, Seoul, South Korea. ✉email: eujinpark@korea.ac.kr

disease is associated with substantial comorbidities and imposes a significant clinical and socioeconomic burden, as patients ultimately require costly kidney replacement therapies (KRTs) such as dialysis or transplants[4,5]. Consequently, the early prediction of disease progression and timely intervention to slow its progression are critical clinical goals, where AI-based predictive models can serve as powerful tools[6–8].

Despite this bright outlook, implementing medical AI models in real-world clinical settings faces significant challenges. The most critical obstacle is the domain-shift problem[9]. This phenomenon occurs when an AI model trained on data from one institution (the source domain) exhibits a substantial degradation in predictive performance when applied to data from another (the target domain). Such performance drops undermine the robustness and trustworthiness of medical AI systems, creating a major technical barrier to their widespread adoption. To address this, recent research has focused on domain generalization[10] and other robust learning approaches designed to enhance model stability across diverse environments.

The challenge of domain shift is particularly pronounced in pediatric medicine. Although children constitute a small proportion of CKD patients, their prolonged disease course and lifelong complications create a pressing need for AI-supported clinical decision-making. This need is contrasted with the reality that pediatric data are inherently scarcer and of lower quality than those of adults[11]. These constraints underscore the critical importance of single-source domain generalization (SSDG), the central focus of the present study.

We aim to address the challenge of developing robust clinical prediction models trained on a single source domain that can maintain performance when applied to unseen target domains.

Related work

SSDG

SSDG has been a particularly active area of research in computer vision[12,13]. The core strategy in this field involves decomposing image data into two components: content and style. Content refers to the "domain-invariant" features, such as an object's essential shape and structure, which remain consistent across domains. By contrast, style encompasses the "domain-variant" features, such as color palette, texture, and lighting, which vary across domains.

Based on this content-style separation, researchers employ strategies that artificially diversify an image's style, thereby training the model to focus only on the content. This has led to the development of successful data-augmentation techniques, such as manipulating an image's frequency characteristics via the Fourier Transform[14–17] or applying deep-learning-based style transfer[18] to render a single image in various artistic styles.

However, these computer-vision-based SSDG approaches have a clear limitation: they cannot be directly applied to tabular datasets. The successful separation of content and style in computer vision relies on prior knowledge of what constitutes stylistic variation (e.g., texture or color). In this case, with tabular data from a single source, we inherently lack the prior knowledge of inter-hospital differences needed to distinguish domain-invariant from domain-variant features. Consequently, existing SSDG methods from computer vision are not suited for tabular data, necessitating the development of robust learning and regularization techniques to achieve domain generalization in this context.

Mixup

Mixup[19] refers to a class of methods that regularize deep networks by creating novel training instances from multiple samples. The foundational method generates virtual samples through linear interpolation of two data points and their corresponding labels. This concept has been extended by other approaches, such as Manifold Mixup[20], which applies interpolation within the network's hidden layers rather than at the input level.

Other popular variations are tailored specifically for computer vision. For instance, CutMix[21] combines samples by cutting a patch from one image and pasting it onto another, with labels mixed proportionally to the patch area. Subsequent methods have refined this patch-based strategy: SaliencyMix[22] leverages saliency detection to ensure the cropped region is informative, whereas ResizeMix[23] simplifies the process by resizing a source image into a small patch.

However, these methods present a fundamental problem in a clinical context. They either combine patient data randomly, without considering clinical plausibility[19,20], or rely on priors and characteristics specific to the vision domain[21–23]. For example, mixing the data of a stable, early-stage patient with that of a late-stage patient on the verge of requiring kidney replacement can create a clinically implausible virtual patient. Such unrealistic data can act as noise during model training and degrade performance, underscoring the need for more context-aware augmentation approaches.

Input masking

Input masking[24], also known as input dropout, is another common technique for enhancing model robustness. This method randomly sets a fraction of the input features to zero (or a baseline value) during training. This procedure prevents the model from becoming overly reliant on any single feature, forcing it to learn a more distributed and robust data representation. However, like standard Mixup, conventional input masking has significant limitations in clinical settings. It typically applies masking with a uniform probability across all features, thereby assuming that feature dropout is random and independent. In practice, missing data in clinical environments often follow specific, non-random patterns. For instance, an entire set of related lab values from a single blood test may be missing concurrently. The failure of standard masking to account for these real-world structured missingness patterns motivates our proposal for a knowledge-guided approach that simulates more realistic missing-data scenarios.

Methods

This section details the proposed methodology and the evaluation benchmark. Figure 1 presents a schematic overview of the entire framework.

Problem setting

Our study's primary goal was to achieve domain generalization (DG) within the challenging SSDG setting. This involved training a robust model f_θ exclusively on a single-source domain $\mathcal{D}_s$ that can generalize effectively to an unseen target domain $\mathcal{D}_t$.

We formally denote the source and target datasets as $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. The core challenge stems from the domain shift, where these domains are sampled from distinct probability distributions ($P_s \neq P_t$). Each data point $(x_i, y_i) \in \mathcal{D}$ comprises a D-dimensional feature vector $x_i \in R^D$ and binary label $y_i \in \{0,1\}$. The model $f_\theta : R^D \rightarrow \{0,1\}$ is trained on $\mathcal{D}_s$ and subsequently evaluated on $\mathcal{D}_t$.

Proposed method 1: similarity-guided mixup

Conventional Mixup research is characterized by two main approaches: random, pixel-wise interpolation and augmentation based on image-specific priors. Random interpolation is inadequate for clinical tabular data, as interpolating patients without regard to their medical status can generate clinically implausible virtual samples. By contrast, augmentation based on image-specific priors relies on visual priors such as saliency or patch area and is fundamentally ill-suited for tabular data.

Our method overcomes these limitations by defining similarity using a clinically meaningful feature set, adapting the augmentation specifically for medical tabular data. We introduce similarity-guided Mixup, a technique that leverages domain knowledge to identify a subset of clinically significant features, $\mathcal{F}_{\text{clinical}} \subset \{1,2,\ldots,D\}$. Augmentation is then performed by selectively mixing patient samples that exhibit high similarity within this specific feature space.

We define a projection function $\varphi : R^D \rightarrow R^{|\mathbb{F}_{\text{clinical}}|}$ that extracts only these clinically meaningful features from a sample vector x.

Distance calculation

First, the distance $d(x_i, x_j)$ between two samples is calculated using the Euclidean norm (L_2 norm) only within this projected feature space:

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = ||\varphi(\boldsymbol{x_i}) - \varphi(\boldsymbol{x_j})||_2 \tag{1}$$

Similarity score

This distance is then converted into a similarity score $S(x_i, x_j)$ via a Gaussian kernel:

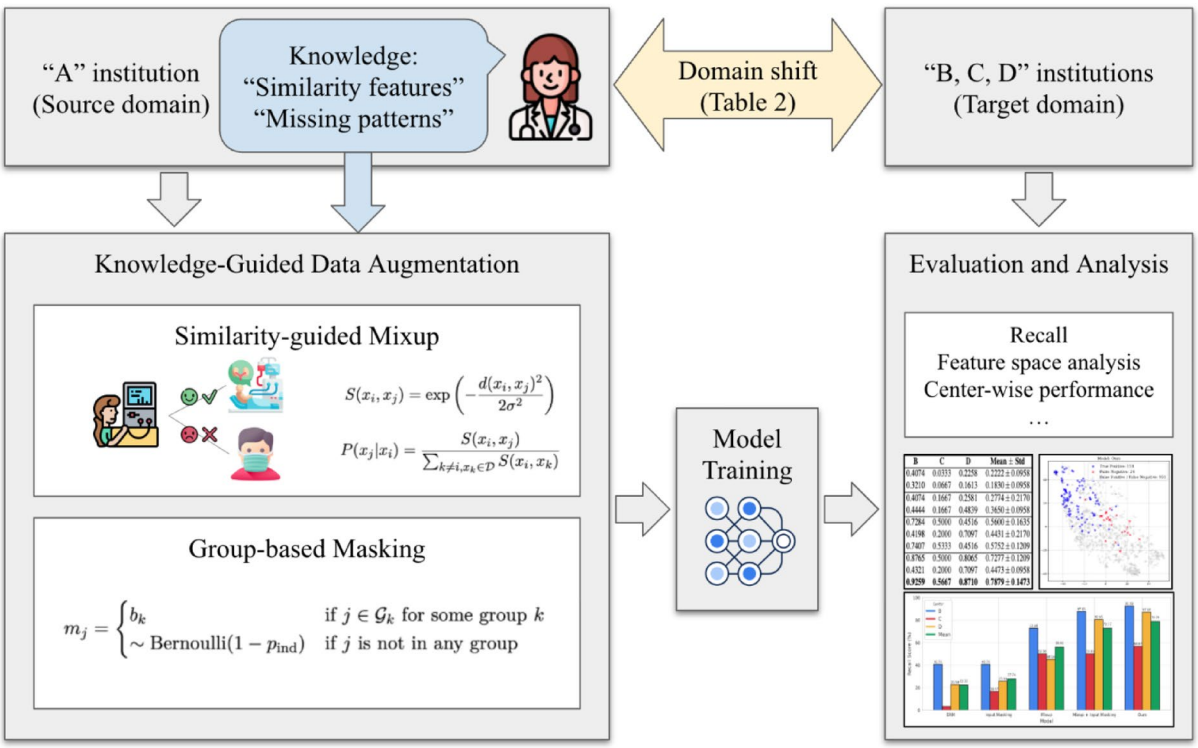

Fig. 1. Overview of the proposed framework and evaluation benchmark. The diagram illustrates the process for evaluating model generalizability under the SSDG setting.

$$S\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \exp\left(-\frac{d\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right)^2}{2\sigma^2}\right) \tag{2}$$

where the hyperparameter σ controls the sensitivity of the similarity conversion.

Sampling probability
The probability $\mathrm{P}(\mathrm{x_j}|\mathrm{x_i})$ of selecting x_j from the dataset D to mix with x_i is derived by normalizing the similarity scores:

$$P\left(x_j|x_i\right) = \frac{S\left(x_i, x_j\right)}{\sum_{k:x_k \in \mathcal{D}} S(x_i, x_k)} \tag{3}$$

This probabilistic sampling ensures that clinically similar samples have a higher likelihood of being mixed, thereby generating more plausible augmented data.

Interpolation
Finally, once a clinically similar pair (x_i, x_j) is selected, the augmented sample (x', y') is generated using the standard Mixup interpolation formulas:

$$x' = \lambda\, x_i + (1 - \lambda)\, x_j \tag{4}$$

$$y' = \lambda\, y_i + (1 - \lambda)\, y_j \tag{5}$$

Where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$.

Proposed method 2: group-based masking

Standard input masking assumes that feature missingness is random and independent. By contrast, our group-based masking incorporates domain knowledge by modeling the correlated missingness patterns common in clinical data and by accounting for more realistic data imperfections to enhance model robustness.

Identifying correlated missing patterns
Through clinical consultation, we identified features that are often missing concurrently (e.g., specific lab panels or paired measurements such as pediatric BP) and organized them into predefined groups $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots\}$.

Masking procedure
The masking procedure generates a binary mask vector $m \in \{0,1\}^D$ based on these groups. For each group G_k, a single binary variable $b_k \sim \mathrm{Bernoulli}(1 - p_{\mathrm{group}})$ is sampled to determine whether the entire group is masked. Individual features j not belonging to any group are masked independently with probability p_{group}. The final mask m_j is determined as follows:

$$m_j = \begin{cases} b_k & \text{if } j \in \mathcal{G}_k \text{ for some group } k \\ \sim \mathrm{Bernoulli}(1 - p_{\mathrm{ind}}) & \text{if } j \text{ is not in any group} \end{cases} \tag{6}$$

The resulting masked vector, $x' = x \otimes m$, contains a mix of correlated (group) and independent (random) masking. This more realistically simulates real-world data imperfections and heterogeneity.

Dataset and evaluation

To validate our proposed methodology, we used data from the KNOW-pedCKD study (ClinicalTrials.gov: NCT02165878; registered June 11, 2014), a multicenter prospective observational cohort of Korean pediatric patients with CKD. Data were collected from teaching hospitals affiliated with seven major pediatric nephrology centers in South Korea; Fig. 2 illustrates the detailed study protocol, including collection periods and patient criteria.

The core research problem was defined as a classification task to predict patient outcomes for an unseen institution (target domain). Specifically, the task was to predict whether a patient's condition would worsen—defined as requiring KRT, death, or a ≥ 20% decline in estimated glomerular filtration rate (eGFR)—within the subsequent year, based on their current visit data.

To analyze SSDG performance, we designated the institution with the largest patient population (A) as the source domain and used the three other large institutions (B, C, and D) as unseen target domains to build our benchmark. Patient records, collected sequentially during annual visits, were segmented into person-visit units. During preprocessing, we standardized continuous features to minimize the impact of scale differences and used one-hot encoding for categorical features to prevent the model from assuming spurious ordinal relationships. Missing data during model training were handled via Multiple imputation by chained equations with an imputer trained on the raw training set. Table 1 presents the statistical discrepancy between the source and target domains, quantified using, Maximum Mean Discrepancy, Wasserstein distance, and correlation distance. Also, Tables 2 and 3 shows a p-value analysis. These confirmed significant differences in feature distributions between the source and target domains, statistically validating the existence of a domain shift.

For model evaluation, we employed a dual-metric strategy. The optimal model checkpoint was selected as the epoch with the highest area under the receiver operating characteristic curve (AUROC) on the validation set, as

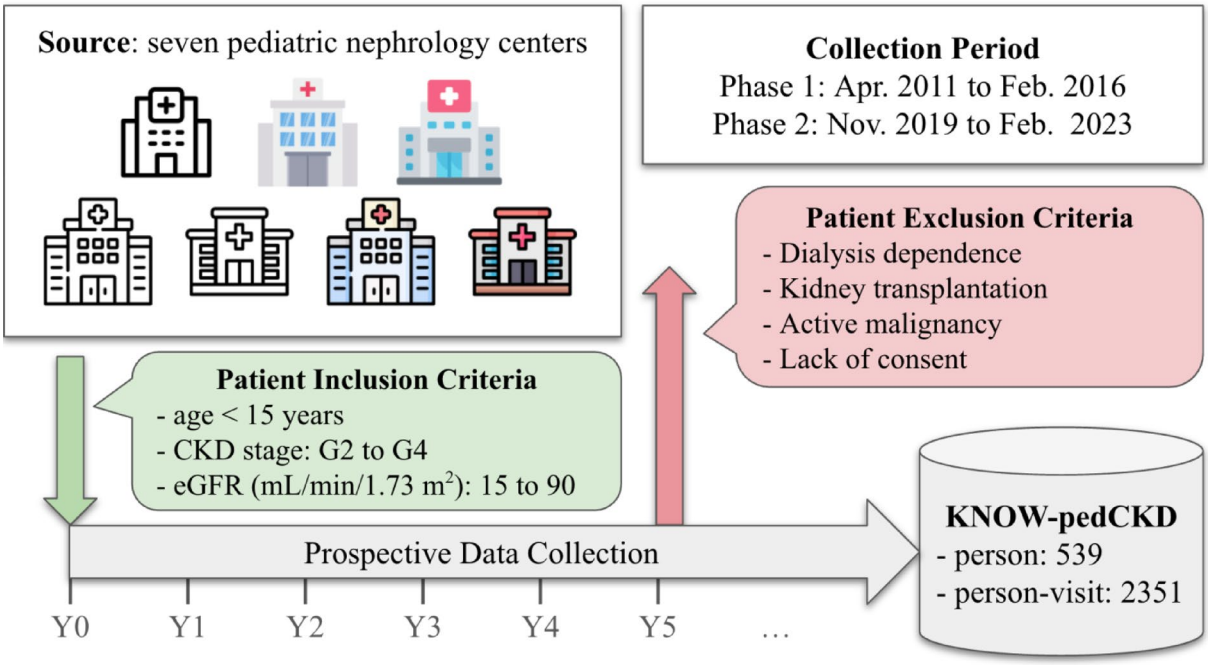

Fig. 2. Overview of the KNOW-pedCKD dataset collection. The diagram illustrates the data-acquisition process for the KNOW-pedCKD cohort, detailing the source institutions, collection period, and the specific patient inclusion and exclusion criteria.

Statistical measures	B (Target 1)	C (Target 2)	D (Target 3)
Maximum mean discrepancy	0.0034	0.0042	0.0043
Wasserstein distance	14.0506	16.1063	49.2632
Correlation distance	0.1323	0.1485	0.1403

Table 1. Statistical distance measures between the source domain (A) and target domains (B, C, and D). We use Maximum Mean Discrepancy, Wasserstein distance, Correlation distance to measure the statistical distance between domains. All domains show significant values across all metrics, confirming that the presence of domain shift. Notably, B has the minimum values across all metrics, indicating that B can be treated as the most closely related domain to A, the source domain.

AUROC provides a comprehensive, threshold-independent measure of overall model discrimination. Moreover, recall was selected for reporting the final performance across methods. This metric was prioritized because, in a clinical setting with critical conditions such as kidney failure, maximizing the detection of true-positive (TP) cases is paramount to minimizing missed intervention opportunities.

Baselines

We compared the performance of the proposed method against that of several baselines. For non-augmentation-based approaches, we included empirical risk minimization (ERM) and invariant risk minimization (IRM)[25]. As IRM is inherently designed for multisource domain generalization, we adapted it to our SSDG setting by creating pseudo-domains. This was accomplished by clustering patients within the SNUH source data, thereby simulating a multisource environment for training. For augmentation-based methods, which served as the non-knowledge-guided baselines, we selected Mixup[19], Manifold Mixup[20], and input masking. We also included traditional imbalanced learning methods, SMOTE[26] and BorderlineSMOTE[27], becuase we use balanced augmentation strategy for Mixup-based approaches.

Implementation details

All data preprocessing techniques, including standardization, one-hot encoding, and MICE imputation[28], were implemented using the scikit-learn package. The employed model architecture was a multi-layer perceptron (MLP). Given the input size, the network was configured with two hidden layers of 32 and 16 nodes, respectively, both using the rectified linear unit activation function. Training was conducted for 2,000 epochs using the ADAM[29] optimizer with a fixed learning rate of 0.0001. Consistent with our evaluation strategy, the final performance was reported for the best model, which was selected based on the most balanced AUROC achieved across all training epochs.

Characteristic	A (Source)	B (Target 1)	C (Target 2)	D (Target 3)
Sex, Male	333 (30.69%)	144 (32.88%)	102 (29.91%)	111 (35.35%)
Age, years	12.85 ± 5.91	12.10 ± 6.35	13.69 ± 5.45	14.43 ± 5.11
eGFR, mL/min/1.73m^2	58.05 ± 29.24	57.39 ± 32.09	61.90 ± 26.32	89.85 ± 29.52
Systolic BP, mmHg	111.36 ± 12.49	113.98 ± 15.36	114.63 ± 11.39	117.76 ± 15.76
Diastolic BP, mmHg	66.65 ± 10.93	68.86 ± 12.32	70.01 ± 9.18	66.42 ± 11.57
Hemoglobin, g/dL	13.14 ± 1.88	12.51 ± 2.06	13.58 ± 1.86	13.20 ± 1.79
Reticulocyte, %	1.48 ± 1.25	1.34 ± 0.87	1.63 ± 0.63	1.25 ± 0.49
Potassium, mmol/L	4.40 ± 0.52	4.28 ± 0.61	4.40 ± 0.43	4.36 ± 0.46
Chloride, mmol/L	105.52 ± 3.52	105.45 ± 3.53	103.21 ± 3.15	104.48 ± 2.89
Calcium, mg/dL	9.60 ± 0.54	9.29 ± 0.63	9.43 ± 0.52	9.43 ± 0.64
Phosphate, mg/dL	4.47 ± 0.81	4.86 ± 4.19	4.29 ± 0.70	4.20 ± 0.87
Albumin, g/dL	4.26 ± 0.42	3.99 ± 0.56	4.36 ± 0.39	4.36 ± 0.49
UPCR, mg/mg	1.19 ± 2.91	2.01 ± 7.68	0.71 ± 0.95	0.72 ± 1.79
Calcium phosphate binder	279 (27.41%)	80 (18.31%)	59 (17.61%)	10 (3.25%)
Iron supplements				
Oral	218 (20.15%)	62 (14.16%)	29 (8.53%)	18 (5.73%)
Intravenous	6 (0.55%)	4 (0.91%)	0 (0.00%)	0 (0.00%)
Intravenous and oral	4 (0.37%)	21 (4.79%)	0 (0.00%)	0 (0.00%)
Not use	854 (78.93%)	351 (80.14%)	311 (91.47%)	296 (94.27%)
ESA				
Epoetin alfa	9 (0.83%)	13 (2.97%)	0 (0.00%)	2 (0.64%)
Epoetin beta	86 (7.96%)	21 (4.79%)	0 (0.00%)	1 (0.32%)
Darbepoetin alfa	13 (1.20%)	3 (0.68%)	0 (0.00%)	3 (0.96%)
CERA	2 (0.19%)	4 (0.91%)	0 (0.00%)	0 (0.00%)
Not use	971 (89.82%)	397 (90.64%)	340 (100.00%)	308 (98.09%)
ACE inhibitor	304 (29.09%)	144 (33.18%)	135 (40.42%)	208 (66.45%)
ARB	384 (36.75%)	60 (13.82%)	30 (8.98%)	91 (29.07%)

Table 2. Demographics and baseline clinical characteristics of the study cohorts. Data are presented for the source domain (A) and the three target domains (B, C, and D). eGFR, estimated glomerular filtration rate; BP, blood pressure; UPCR, urine protein to creatinine ratio; ESA, Erythropoietin Stimulating Agents; CERA, Continuous Erythropoietin Receptor Activator; ACE, Angiotensin-Converting Enzyme; ARB, Angiotensin Receptor Blocker.

For similarity-guided Mixup, clinical similarity between patients was defined using features identified as critical in the widely used Pediatric Estimated Time to KRT Calculator. This web-based clinical prediction tool was developed and adapted from the Chronic Kidney Disease in Children (CKiD) study, a large, prospective, multicenter cohort established in the United States and Canada[30]. This calculator predicts KRT timing based on key clinical parameters, which we used for our similarity metric: urine protein-to-creatinine ratio (UPCR), blood pressure, eGFR, hemoglobin level, serum albumin, chloride, and bicarbonate concentrations[31].

For group-based masking, we established predefined feature groups based on clinical correlations and practice: "Diastolic BP" and "Systolic BP" were grouped because they are measured concurrently, and angiotensin II receptor blockers (ARBs) and angiotensin-converting enzyme (ACE) inhibitors were grouped because they belong to the same drug class.

For hyperparameters, in Similarity-based Mixup, we set $\sigma = 1, \lambda \in \{0.2, 0.3, 0.4\}$. For Group-based Masking, we utilized $p_{ind} \in \{0.4, 0.5, 0.6\}$, and applied same values to p_{group} rather than exploring all combinations. To ensure a fair comparison, identical hyperparameters were applied to both the Mixup and Masking baselines. Furthermore, for our proposed method as well as the Mixup-based and SMOTE-based baselines, a balanced augmentation strategy was consistently employed across both classes.

Ethical considerations

All procedures involving human participants were performed in accordance with relevant guidelines and regulations. This study was reviewed and approved by the Institutional Review Board of Seoul National University Hospital (No. H-1906068-1041). Informed consent was obtained from participants or their legal guardians, as appropriate, based on age. Development and reporting of the prediction models adhered to the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guideline to ensure transparent and reproducible methodology.

Characteristic, unit	A (Source)	B (Target 1)	C (Target 2)	D (Target 3)
Count (% of patients)	1085 (15.48%)	438 (18.49%)	341 (8.80%)	314 (9.87%)
Sex	-	0.41	0.79	0.12
Age, years	-	0.02	0.04	0
eGFR, mL/min/1.73m^2	-	0.35	0.01	0
Systolic BP, mmHg	-	0.01	0	0
Diastolic BP, mmHg	-	0	0	0.31
Hemoglobin, g/dL	-	0	0	0.28
Reticulocyte, %	-	0.04	0	0.04
Potassium, mmol/L	-	0	0.53	0.05
Chloride, mmol/L	-	0.66	0	0
Calcium, mg/dL	-	0	0	0
Phosphate, mg/dL	-	0.01	0	0
Albumin, g/dL	-	0	0	0
UPCR, mg/mg	-	0	0.07	0
Calcium phosphate binder	-	0	0	0
Iron supplements	-	0.33	0	0
ESA	-	0.70	0	0
ACE inhibitor	-	0.12	0	0
ARB	-	0	0	0.01

Table 3. Statistical comparison of feature distributions between the source and target domains. P-values are reported for continuous variables, whereas chi-square test p-values are used for categorical variables. The results confirm a significant difference in distribution for most features, validating the presence of domain shift. eGFR, estimated glomerular filtration rate; BP, blood pressure; UPCR, urine protein to creatinine ratio; ESA, Erythropoietin Stimulating Agents; ACE, angiotensin-Converting Enzyme; ARB, Angiotensin Receptor Blocker.

MLP	Method	B	C	D	Mean ± STD
W/o augmentation	ERM	0.4074	0.0333	0.2258	0.2222 ± 0.0958
	IRM	0.3210	0.0677	0.1613	0.1830 ± 0.0958
Masking only	Input masking	0.4074	0.1667	0.2581	0.2774 ± 0.2170
	Group-based masking	0.4444	0.1667	0.4839	0.3650 ± 0.0958
Mixup only	Mixup	0.7284	0.5000	0.4516	0.5600 ± 0.1635
	Manifold Mixup	0.4198	0.2000	0.7097	0.4431 ± 0.2170
	Similarity-guided Mixup	0.7407	0.5333	0.4516	0.5752 ± 0.1209
SMOTE-based	SMOTE	0.7836	0.7621	0.6901	0.6354 ± 0.0490
	BorderlineSMOTE	0.7788	0.7593	0.6197	0.5674 ± 0.0868
Mixup + Masking	Mixup + Input masking	0.8765	0.5000	0.8065	0.7277 ± 0.1209
	Manifold Mixup + Masking	0.4321	0.2000	0.7097	0.4473 ± 0.0958
	Ours	**0.9259**	**0.5667**	**0.8710**	**0.7879 ± 0.1473**

Table 4. Recall performance comparison in the SSDG setting. The table displays center-wise recall scores for each method across the target domains (B, C, and D), along with the mean (± standard deviation) across all centers. The proposed method achieves the highest mean recall compared to all baselines.

Results

Overall performance comparison

Table 4 presents the Recall values for the various methods, demonstrating that the proposed method yields significant performance improvements. Our method achieved a mean recall of 0.7879, outperforming the best-performing baseline (Mixup + Masking) by a substantial margin of 6.20% points. This result highlights that our approach is particularly effective at identifying patients at risk of deterioration across unseen domains.

Ablation of proposed components

Table 4 also details an ablation study for each component of our method. Compared to the ERM baseline, our knowledge-guided components individually outperformed their conventional counterparts. Group-based

Mixup	Masking	B	C	D	Mean ± STD
Random-feature-based	Group-based (ours)	0.9012	0.5333	0.8710	0.7685 ± 0.1693
Similarity-guided (ours)	Random-group-based	0.8765	0.5000	0.8387	0.7384 ± 0.1693
Similarity-guided (ours)	**Group-based (ours)**	**0.9259**	**0.5667**	**0.8710**	**0.7879 ± 0.1473**

Table 5. Effectiveness of similarity-guided mixup and group-based masking. The performance degradation observed in Random-feature-based Mixup and Random-group-based Masking highlights the importance of incorporating clinical priors and distributional similarity.

Method	A-based	B-based	C-based	D-based
Vanilla	0.2222 ± 0.0958	0.0334 ± 0.0323	0.1776 ± 0.0953	0.1111 ± 0.0135
Mixup	0.5600 ± 0.1635	0.5768 ± 0.1760	0.6088 ± 0.2113	0.5971 ± 0.1263
Mixup + Input masking	0.7277 ± 0.1209	0.7946 ± 0.1224	0.8992 ± 0.0871	0.7795 ± 0.1725
Ours	**0.7879 ± 0.1473**	**0.8868 ± 0.1047**	**0.9396 ± 0.0642**	**0.8574 ± 0.1000**

Table 6. Robustness evaluation across various source domain configuration. This table presents the performance of our proposed SSDG framework when rotating the source domain among B, C, and D. The consistent outperformance of our method over the baselines in every setting demonstrates its robustness to source domain variation.

masking achieved a greater performance improvement than standard input masking, and similarity-guided Mixup likewise demonstrated a larger gain than standard Mixup.

Effectiveness of each components

To validate the necessity of Similarity-guided Mixup, we compared it against a Random-feature-based Mixup baseline. The random variant selects an equivalent number of features, excluding those used in our method, to calculate distances between samples. Similarly, Group-based Masking was compared to a random-group version where features were grouped without any medical logic. All experiments were conducted three times to obtain the best scores of random counterparts. As shown in Table 5, our proposed method performed better than the random baselines, suggesting that leveraging clinical knowledge is essential for the model to understand complex patient data.

Source domain robustness

To further evaluate the robustness and generalizability of our approach, we conducted additional experiments by rotating the source domain among B, C, and D. This setup ensures that the performance gains of Similarity-guided Mixup and Group-based Masking are not dependent on a specific source distribution but remain consistent across diverse data environments. As shown in Table 6, our method outperformed the baselines in every source domain variations.

Model robustness

To evaluate the model-agnostic effectiveness of our approach, we extended our experiments beyond the MLP to include gradient-boosted decision trees, such as XGBoost and CatBoost[32,33]. The consistent performance gains across these diverse architectures demonstrate the model robustness of our proposed SSDG framework, as shown in Table 7.

Feature-space analysis

To qualitatively analyze how our method improves feature representation, we visualized the test set's latent space using t-distributed stochastic neighbor embedding (t-SNE)[34] (Fig. 3). These visualizations demonstrate that our framework learns a more structured and discriminative feature space than the baselines. Compared with methods such as ERM, our model's TPs (blue dots) and false negatives (FNs, red crosses) are more clearly clustered and separated. Notably, our approach achieves a dramatic reduction in FNs (red "x" marks), misclassifying only 24 positive cases. This is a considerable improvement over the 93 FNs observed with ERM and 54 with standard Mixup. This reduction confirms that our dual knowledge-guided augmentation framework compels the model to learn a robust decision boundary that better captures the data's inherent structure, thereby maximizing the identification of high-risk patients across unseen target domains.

Center-wise analysis

We observe that performance generally improved across all target hospitals, consistent with the overall mean performance gain. Notably, performance in Center C showed a substantial gain upon the incorporation of Mixup, suggesting that Mixup's augmentation, which increases the number of samples in the minority class, was a significant contributing factor to performance in that specific domain.

Architecture	Method	B	C	D	Mean ± STD
XGBoost	Vanilla	0.7933	0.7917	0.4085	0.3617 ± 0.2217
	Mixup	0.7811	0.7778	0.5352	0.4834 ± 0.1410
	Mixup + Input masking	0.7831	0.7696	0.6268	0.5844 ± 0.0866
	Ours	**0.7983**	**0.7912**	**0.6268**	**0.5917 ± 0.0970**
CatBoost	Vanilla	0.8069	0.8016	0.4225	0.3623 ± 0.2204
	Mixup	0.8053	0.7991	0.5634	0.5131 ± 0.1379
	Mixup + Input masking	0.7900	0.7728	0.5704	0.5102 ± 0.1221
	Ours	**0.7921**	**0.7839**	**0.5845**	**0.5324 ± 0.2204**

Table 7. Model-agnostic evaluation. The performance is compared across different architectures, including XGBoost and CatBoost, to demonstrate the universal effectiveness of our approach. Notably, our method consistently achieves superior mean performance compared to vanilla and other augmentation baselines across all evaluated architectures.

Discussion

This study addressed the critical challenge of domain generalization in clinical prediction, where AI models trained at one institution often fail when applied to unseen target domains. We proposed a dual knowledge-guided data-augmentation framework, grounded in an explicit clinical context, to address this problem in the challenging single-source setting. Our experimental results validated this approach, demonstrating that our method substantially outperforms standard baselines—including ERM, IRM, Mixup, and input masking—particularly on the clinically vital recall metric.

The framework's success stems from embedding clinical context directly into the augmentation process. Conventional Mixup generates noise by indiscriminately interpolating clinically disparate patients, creating unrealistic data that can degrade performance. By contrast, our similarity-guided Mixup avoids this pitfall by selectively interpolating only clinically similar patients. This preserves the data's intrinsic structure and functions as an effective regularizer, generating plausible virtual samples that enhance data diversity.

Similarly, our group-based masking simulates realistic data imperfections—such as the concurrent absence of related laboratory values—a correlated pattern that standard uniform masking ignores. This forces the model to learn more stable representations that are resilient to the structured missingness patterns encountered in real-world data.

This knowledge-guided strategy is in sharp contrast to baseline models such as IRM, whose performance was notably limited. We attribute this failure to the inadequacy of using pseudo-domains. Creating clusters from a single source cannot replicate the true, complex heterogeneity in patient demographics, measurement protocols, and data recording practices across different institutions. This finding underscores that effective domain generalization requires simulating true inter-domain diversity, not merely enforcing invariance within a single, limited source.

Our work offers two primary contributions. First, we empirically demonstrate the necessity of embedding domain knowledge into AI models for tabular medical data, a domain where generic augmentation techniques from other fields often underperform. Unlike images, EMR tabular data lack spatial structure and involve complex nonlinear interactions, causing traditional Mixup to produce clinically implausible samples. To address this limitation, our method incorporates clinical knowledge to ensure the generation of medically valid synthetic data. Second, we present a practical and effective methodology for the single-source setting that accurately reflects the data accessibility constraints common in healthcare AI research. These knowledge-driven regularizations meaningfully enhance both model robustness and clinical interpretability.

However, this study had several limitations, which must be acknowledged. Our validation was confined to a specific pediatric CKD dataset (KNOW-pedCKD), and the identification of "clinically significant features" relied on expert knowledge, which could introduce subjectivity. We focus on an interpretable, knowledge-based simple approach for data augmentation. Consequently, resource-intensive Transformer-based generative methods were excluded from this study. And as shown in Table 8, although the AUROC of our method is relatively lower than SMOTE-based approaches, it exhibits a robust Recall performance in Table 4. This trade-off is clinically acceptable as it enhances the model's sensitivity to high-risk patients, fulfilling the primary requirement of a clinical decision support system. While we used the challenging SSDG setting as a rigorous testbed, our primary contribution is this novel regularization framework rather than a definitive solution to the SSDG problem itself. Therefore, future research should validate the framework's generalizability across diverse clinical datasets and adult populations. We also see value in exploring data-driven approaches for automated feature discovery and extending this methodology to multisource settings. By integrating deep domain knowledge with advanced augmentation, this study provides a promising pathway toward developing medical AI models that are sufficiently robust and trustworthy for deployment across heterogeneous clinical environments.

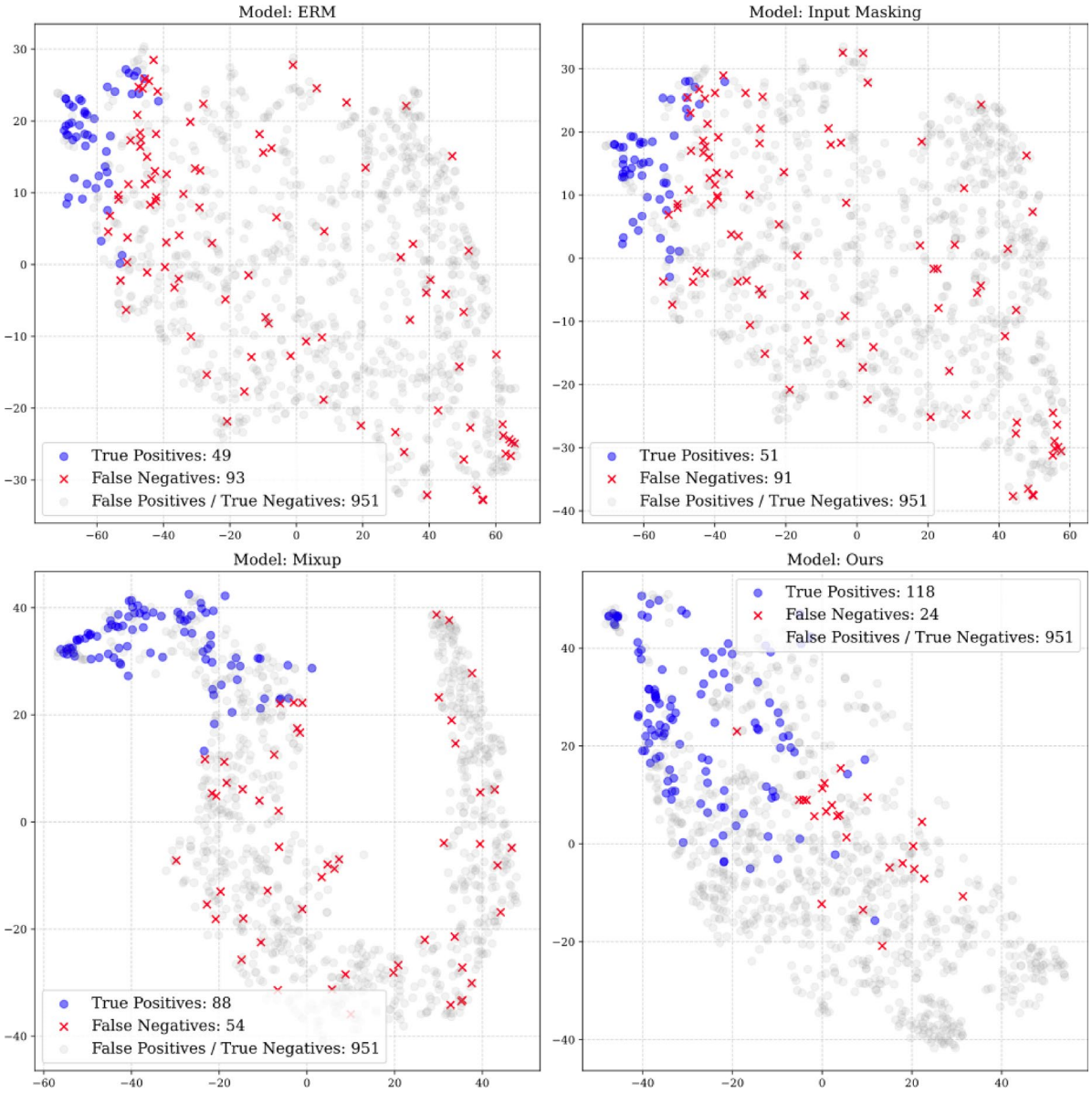

Fig. 3. Feature-space analysis via t-SNE visualization. The plots visualize the latent space of the test set for different models, highlighting TPs (blue dots) and FNs (red crosses). The negative class is faded for clarity. Our method (bottom right) yields a more structured feature space with improved class separation, resulting in a significant reduction in FNs.

MLP	Method	B	C	D	Mean ± STD
W/o augmentation	ERM	0.7606	0.7511	0.7165	0.7427 ± 0.0189
	IRM	0.7609	0.7592	0.7158	0.7453 ± 0.0208
Masking only	Input masking	0.6875	0.5704	0.6587	0.6389 ± 0.0498
	Group-based masking	0.6703	0.5175	0.6365	0.6081 ± 0.0655
Mixup only	Mixup	0.7671	0.7699	0.7390	0.7586 ± 0.0140
	Manifold Mixup	0.4316	0.2265	0.4713	0.3765 ± 0.1073
	Similarity-guided Mixup	0.7678	0.774	0.7344	0.7587 ± 0.0174
SMOTE-based	SMOTE	0.7749	0.766	0.7455	0.7621 ± 0.0855
	BorderlineSMOTE	0.768	0.7702	0.7395	0.7593 ± 0.1075
Mixup + Masking	Mixup + Input masking	0.7361	0.5393	0.6995	0.6583 ± 0.0898
	Manifold Mixup + Masking	0.4331	0.2269	0.4718	0.3773 ± 0.0123
	Ours	0.7248	0.5205	0.6929	0.6461 ± 0.0140

Table 8. AUROC values of Table 4. This table summarizes the AUROC performance of Table 4. While our method exhibits a competitive AUROC, the results should be interpreted alongside its superior Recall performance, which prioritizes the identification of high-risk patient trajectories in imbalanced clinical data.

Data availability

The analytic datasets presented in this study are available from the corresponding author upon reasonable request.

Received: 3 November 2025; Accepted: 26 March 2026
Published online: 03 April 2026

References

1. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
2. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).
3. Sutton, R. T. et al. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digit. Med.* **3**, 17 (2020).
4. Ng, D. K. et al. Incidence of initial renal replacement therapy over the course of kidney disease in children. *Am. J. Epidemiol.* **188**, 2156–2164 (2019).
5. Xie, Y. et al. Analysis of the global burden of disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. *Kidney Int.* **94**, 567–581 (2018).
6. Islam, M. A., Majumder, M. Z. H. & Hussein, M. A. Chronic kidney disease prediction based on machine learning algorithms. *J. Pathol. Inf.* **14**, 100189 (2023).
7. Ghosh, S. K. & Khandoker, A. H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci. Rep.* **14**, 3687 (2024).
8. Moon, S. et al. Development of a prediction tool for kidney function decline in children with chronic kidney disease. *Kidney Res. Clin. Pract.* https://doi.org/10.23876/j.krcp.25.004 (2025).
9. Kanakasabapathy, M. K. et al. Adaptive adversarial neural networks for the analysis of lossy and domain-shifted datasets of medical images. *Nat. Biomed. Eng.* **5**, 571–585 (2021).
10. Guo, L. L. et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci. Rep.* **12**, 2726 (2022).
11. Harada, R., Hamasaki, Y., Okuda, Y., Hamada, R. & Ishikura, K. Epidemiology of pediatric chronic kidney disease/kidney failure: Learning from registries and cohort studies. *Pediatr. Nephrol.* **37**, 1215–1229 (2022).
12. Xu, Q. et al. Simde: A simple domain expansion approach for single-source domain generalization In CVPR workshops. 4798–4808 (IEEE, 2023).
13. Cugu, I., Mancini, M., Chen, Y. & Akata, Z. Attention consistency on visual corruptions for single-source domain generalization in CVPR workshops. 4164–4173 (IEEE, 2022).
14. Xu, Q., Zhang, R., Zhang, Y., Wang, Y. & Tian, Q. A fourier-based framework for domain generalization in CVPR (2021).
15. Zhao, H. et al. Morestyle: Relax low-frequency constraint of fourier-based image reconstruction in generalizable medical image segmentation in MICCAI (ed Linguraru, M. G.) 434–444 (Springer Nature Switzerland, 2024).
16. Liu, C., Cao, Y., Su, X. & Zhu, H. Universal frequency domain perturbation for single-source domain generalization in *Proc. 32nd ACM international conference on multimedia* 6250–6259 (ACM, 2024).
17. Huang, J., Guan, D., Xiao, A. & Lu, S. Fsdr. Frequency space domain randomization for domain generalization in CVPR (2021).
18. Zhou, K., Yang, Y., Qiao, Y. & Xiang, T. Domain generalization with mixstyle in ICLR (2021).
19. Zhang, H., Cissé, M. & Dauphin, Y. N. & Lopez-Paz, D. mixup: beyond empirical risk minimization in ICLR (2018).
20. Verma, V. et al. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, (2019).
21. Yun, S. et al. Cutmix: Regularization strategy to train strong classifiers with localizable features in ICCV (2019).
22. Uddin, A. F. M., Monira, M., Shin, W., Chung, T. & Bae, S. H. Saliencymix: A saliency guided data augmentation strategy for better regularization in ICLR (2021).
23. Qin, J. et al. Resizemix: Mixing data with preserved object information and true labels. *arXiv* (2020).
24. Balasubramanian, S. & Feizi, S. Towards improved input masking for convolutional neural networks in ICCV 1855–1865 (IEEE, 2023).
25. Arjovsky, M., Bottou, L. & Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. *arXiv* (2019).
26. Chawla, N. V. et al. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2011).
27. Han, H., Wang, W. Y. & Mao, B. H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Int. Conf. Intell. Comput.* 878–887 (2005).

28. Van Buuren, S. & Groothuis-Oudshoorn, K. Mice: multivariate imputation by chained equations in r. *J. Stat. Softw.* **45**, 1–67 (2011).
29. Kingma, D. P. & Ba, J. A. A method for stochastic optimization in ICLR (2015).
30. Wong, C. J., Moxey-Mims, M., Jerry-Fluker, J., Warady, B. A. & Furth, S. L. Ckid (ckd in children) prospective cohort study: A review of current findings. *Am. J. Kidney Dis.* **60**, 1002–1011 (2012).
31. Ng, D. K. et al. Development of an adaptive clinical web-based prediction tool for kidney replacement therapy in children with chronic kidney disease. *Kidney Int.* **104**, 985–994 (2023).
32. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 785–794 (2016).
33. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features in NeurIPS (2018).
34. Maaten, L. V. D. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9Nov**, 2579–2605 (2008).

Acknowledgements
We gratefully acknowledge the valuable assistance of Heejung Ahn and Sungkyung Kim at the Medical Research Collaborating Center, Biomedical Research Institute of Seoul National University Hospital, for their expert support in data management.

Author contributions
S.M. and E.P. conceptualized the study. S.M. conducted the methodology and wrote the original draft. E.P. reviewed and edited the manuscript and supervised the work. All authors contributed to data curation, investigation, and resources. All authors have read and approved the final version of the manuscript.

Funding
This work was supported by the research program of the National Institute of Health (NIH) under research project No. 2025E110100. Additional support was provided by a grant from Korea University Medicine (No. K2427191).

Declarations

Competing interests
The authors declare no competing interests.

Ethical consent
This study was reviewed and approved by the Institutional Review Board of Seoul National University Hospital (No. H-1906068-1041).

Informed consent
Informed consent was obtained from participants or their guardians, as appropriate, based on their age.

Additional information
Correspondence and requests for materials should be addressed to E.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.